Detecting Adversarial Examples via Undercover Attack

Qifei Zhou
School of Software and Microelectronics
Peking University
Beijing 100871, China
qifeizhou@pku.edu.cn

Weiping Li
School of Software and Microelectronics
Peking University
Beijing 100871, China
wpli@ss.pku.edu.cn

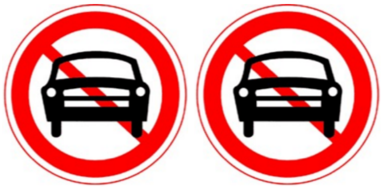
Tong Mo
School of Software and Microelectronics
Peking University
Beijing 100871, China
motong@ss.pku.edu.cn

Yue Wu
Logistics Engineering
University of Science and Technology Beijing
Beijing 100083, China
869216739@qq.com

Abstract

1 Introduction

Deep nerual networks (DNNS) are vulneralbe to adverisarial examples [14, 7, 11], imperceptible modifications to the original inputs of a DNN classifier can cause the model to produce incorrect labels. This problem is especially important in safety-critical applications such as self-driving cars. To illustrate how adversarial samples make a system based on DNNs vulnerable, consider the following images 1:



(a) prohibition sign (b) fire sign

Figure 1. (a) is the original image of a prohibition sign

To humans, these two images look the same: we identify each of them as a prohibition sign. The image on the left is indeed an ordinary image of a prohibition sign. We produced the image on the right by adding a very small perturbation that forces a particular DNN to classify it as a fire sign. The work or Kurakin et al. [8] showed that these transformations are effective in the physical world. Here, someone with ulterior motives can make self-driving cars behave dangerously taking advantage of the vulnerability of DNNs.

There are quite a few methods have been proposed to fool the model, such as L-BFGS [14], FGSM [7], BIM [8], JSMA [12] and C&W [2]. This problem has aroused great concern in the academic world. Many researchers has being trying to explain adversarial examples and find new ways to defend aganist these attack methods. There are numerous defense techniques including image compression

Preprint. Work in progress.

or filtering [4, 15], defensive distillation [13] and many defenses summarized as Gradient masking [11] or Obfuscated Gradients [1]. Unfortunately, none of these defenses is yet completely satisfactory, they can generally be evaded by stronger attacks wholly or partially. The most popular defense in current research papers is probably adversarial training proposed by Goodfellow et al. [7] which also helps us a lot in *Undercover Attack*.

Due to the challenge of defenses, many recent work has turned to detect adversarial examples. Feinman et al. [6] showed that Kernel Density estimates and Bayesian Uncertainty estimates can detect points lie in low-confidence regions of the input space. Pang et al. [10] proposed a new loss function named Reverse Cross-entropy which can improve the performance of KD and BU. Howerver, KD and BU have been debeated by C&W in [3] who showed that incorporating kernel density into the objective function makes detection substantially more difficult. Ma et al. [9] proposed Local Intrinsic Dimensionality which is claimed robust to C&W and achieves quite execellent performance on various attacks both in black-box and white-box. KD, BU and LID all rely on knowledge of the attack mechanism. They need to train on adversarial examples generated by these attack methods which can't cope with new attacks. Inspired by the use of biometric and cryptograpic signatures, Dathathri et al. [5] proposed NeuralFingerprinting (NeuralFP). They trained the model with NeuralFP which achieved near perfect AUC-scores aganist black-box attacks. While, just like in cryptography, the model can't guarantee its effectiveness if the NeuralFP is exposed to us.

We have been working hard to study and explain the vulnerability of normal samples all the times. However, we didn't realize that adversarial samples are more vulnerable than normal ones. It may be not easy to sucessfully attack a benign example, and we even can't find any adversarial examples within the ϵ norm ball of the benign example as long as ϵ is small enough. Nevertheless, there must be some small modifications which can attack an adversarial example sucessfully. Because the adversarial example is generated from a benign example, we can find at least one way which is rolling back to the benign one. This is a successful attack for the adversarial example (In our paper, a successful attack is an attack which changes the model's prediction, not necessary the true lable. Especially for an adversarial input, its prediction is not exactly the true class). In fact, our experiments show that adversarial examples is far more vulnerable than normal examples. We design *Undercover Attack* to detect adversarial examples based on this property. In particular, our key contributions are:

- We propose *Undercover Attack* to detect adversarial examples. *Undercover Attack* is a novel idea which defends by attacks. We discuss how *Undercover Attack* can distinguish adversarial samples, and show that it is easier to attack adversarial samples successfully than normal samples through experiments.
- Our study reveals that even very simple attack mechanism, like FGSM can attack adversarial samples with high success rate. We train our model on the FGSM attack strategy, after training, our model is robust to FGSM attack on normal samples. However, if we generate adversarial examples through stronger attack mechanisms and then utilize FGSM to attack these adversarial examples, we can always succeed.
- *Undercover Attack* does not rely on knowledge of the attack mechanism. Our framework does not require additional detectors, networks or parameters. We only need litte more Computing resource.
- We emperically show that the performance of *Undercover Attack* is robust to unknown attack methods. In the same difficult scenario: unknown attack and white-box setting, *Undercover Attack* achieves state-of-the-art AUC-scores with an average performace of 97% on various attacks on both MNIST and CIFAR10 datasets.

Acknowledgments

We would like to thank Zhilong Hong and Haochen Li for helpful discussions and their useful advice on the topic.

References

[1] Anish Athalye, Nicholas Carlini, and David A Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *international conference on machine learning*, pages 274–283, 2018.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. 2016.

[3] Nicholas Carlini and David A Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv: Learning*, pages 3–14, 2017.

[4] Nilaksh Das, Madhuri Shanbhogue, Shangtse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv: Computer Vision and Pattern Recognition*, 2017.

[5] Sumanth Dathathri, Stephan Zheng, Richard M Murray, and Yisong Yue. Detecting adversarial examples via neural fingerprinting. *arXiv: Learning*, 2018.

[6] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv: Machine Learning*, 2017.

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *Computer Science*, 2014.

[8] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. 2016.

[9] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi N R Wijewickrema, Grant Schoenebeck, Michael E Houle, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *international conference on learning representations*, 2018.

[10] Tianyu Pang, Chao Du, Yinpeng Dong, and Jun Zhu. Towards robust detection of adversarial examples. *neural information processing systems*, pages 4584–4594, 2018.

[11] Nicolas Papernot, Patrick Mcdaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. 2016.

[12] Nicolas Papernot, Patrick D Mcdaniel, Somesh Jha, Matthew Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *ieee european symposium on security and privacy*, pages 372–387, 2016.

[13] Nicolas Papernot, Patrick D Mcdaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *ieee symposium on security and privacy*, pages 582–597, 2016.

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *Computer Science*, 2013.

[15] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *network and distributed system security symposium*, 2018.